**Capstone Project-4 Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| *Contributor’s Role:*  **Sandeep Kumar Maurya**   * Introduction * Data cleaning * Null Value Treatment * Exploratory Data Analysis * Univariate * Data cleaning & pre-processing for clustering * Encoding the categorical data * K means clustering * Hierarchical Clustering * Silhouette analysis * Conclusion     **Ansh Bhatnagar**   * Introduction * Data cleaning * Null Value Treatment * Exploratory Data Analysis * Univariate * Data cleaning & pre-processing for clustering * Encoding the categorical data * K means clustering * Hierarchical Clustering * Silhouette analysis * Recommendation System * Conclusion |
| **Please paste the GitHub Repo link.** |
| Ansh Bhatnagar:- <https://github.com/AnshRockstar/Netflix-Tv-show-and-Movie-Clustering>  Sandeep Kumar Maurya:- |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches, and your conclusions. (200-400 words)** |
| PROBLEM  This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from flixable which is a third-party Netflix search engine.  In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service’s number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled. It will be interesting to explore what all other insights can be obtained from the same dataset. Integrating, this dataset with other external datasets  such as IMDB ratings, rotten tomatoes can also provide many interesting findings.  APPROACH  Initially, in the 1st step imported the data set to carry out the analysis over the data set to comprehend the details of available data and Checked for Null values and treated them. Here, we found more than 30% null values in the director's column. Then, we take appropriate action for null values according to the circumstances.  ● Analyzing all the variables of the data set and identifying the solution for given tasks.  ● We perform EDA and Data Visualization on our dataset. Here, we found that the proportion of tv shows in Netflix content is very less as compared to the movies. We can observe that the majority of Netflix material is intended for adults. There is very little content available for teens and kids.  **word cloud that from two topics one is about documentaries and other is about kids shows, therefore, topic modelling successfully classifies movies and shows into genre**  ● Performed hypothesis testing to get the insights on duration of movies and content with respect to different variables.  ● After doing feature engineering and finding the number of clusters, we used the k-means algorithm and then checked the model performance using Silhouette’s coefficient, to identify the best fit Model.  The number of movies on Netflix is growing  significantly faster than the number of TV shows. Because of covid-19, there is a  significant drop in the number of movies and television episodes produced after 2019.  Because of covid-19, there is a significant drop in the number of movies and  television episodes produced after 2019.  Performed hypothesis testing to get the insights on duration of movies and content with respect to different variables. We perform the K-Means clustering on our dataset. Here, we find the optimal value of k is 25. But if we want to recommend some movies and tv shows then k=25 is not good, so we did PCA then we take the value of k as 34. The silhouette score for k=25 is 0.027217317155321205. which is a bad score but after PCA Silhouette score is 0.34895060389063276 that is good.  Also performed recommender system with cosine similarities.  CONCLUSION  We have done null value treatment, feature engineering, and EDA since loading the dataset then completed assigned tasks.  Concluded that Netflix is increasingly focusing on movies rather than TV shows, especially after 2014.  Among different types of content available in different countries, content TV-MA is available in most countries. This could be because it shows that it is just for adult audiences, and the Netflix audience enjoys content like this.  We have also explained different clusters based on their content; Defined clusters and enforced the K-means clustering algorithm and cluster number nine has the most clusters; we have also plotted a scatter plot in which we may interact with similar content about that cluster. |